Rare Class Classification by Support Vector Machine

He He — Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China

Ali Ghodsi — Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada

Abstract

The problem of classification on highly imbalanced datasets has been studied extensively in the literature. Most classifiers show significant deterioration in performance when dealing with skewed datasets. In this paper, we first examine the underlying reasons for SVM's deterioration on imbalanced datasets. We then propose two modifications for the soft margin SVM, where we change or add constraints to the optimization problem. The proposed methods are compared with regular SVM, cost-sensitive SVM and two re-sampling methods. Our experimental results demonstrate that this constrained SVM can consistently outperform the other associated methods.

Introduction

In the literature, there are two major groups of methods to address the imbalance problem.

- ***Date level***: oversampling the minority class or undersampling the majority class, e.g. SMOTE [1].
- ***Algorithmic level***: cost-sensitive learning

2C-SVM [1] is the cost-sensitive version of SVM. It essentially reweights the examples to make the error from the rare class more obvious to the classifier.

Primal:

$$\min_{\beta,\beta_0,\xi_i} \quad \frac{1}{2}\beta^T\beta + \gamma C\sum_{i\in I^+}\xi_i + (1-\gamma)C\sum_{i\in I^-}\xi_i$$

$$s.t. \quad y_i(\beta^T x_i + \beta_0) \geq 1-\xi_i$$

$$\xi_i \geq 0$$

* γ is usually set to be the ratio of negative points over positive points

Dual:

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) - \sum_{i=1}^{n}\alpha_i$$

$$s.t. \quad \sum_{i=1}^{n}\alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq \gamma C \quad \text{for} \quad i \in I^+$$

$$0 \leq \alpha_i \leq (1-\gamma)C \quad \text{for} \quad i \in I^-$$

SVM on Imbalanced Datasets

The fact that the SVM solution only depends on a few support vectors makes it relatively robust to noise and moderate imbalance. However, when the datasets is severely imbalanced, SVM shows deterioration.

Experiment

From Fig.1 to Fig.3, as the imbalance becomes more severe, we can conclude that:

- the number of support vectors in the positive class becomes less than the number in the negative class;
- examples from the positive class tend to reside farther from the real boundary than those from the negative class;
- the predicted decision boundary is pressed towards the rare class.

Analysis

- ***Minor Imbalance***

$\sum_{i=1}^{n}\alpha_i y_i = 0$ If there are more negative samples (y_i = -1) than positive samples (y_i = 1), then the positive class will have higher weights (α_i).

$y = sign(\sum_{i=1}^{n}\alpha_i y_i K(x,x_i) + \beta_0)$ Higher weights increase the influence of the minority class, which automatically rebalances the skewed dataset.

- ***Severe Imbalance***

$y_i(\beta^T x_i + \beta_0) = 1-\xi_i < 1$ This is derived from the complementary slackness, indicating that the point has entered the margin or even crossed the decision boundary.

$\alpha_i = C$ When support vectors of the negative class are much more than those of the positive class, α_i is made to equal the maximum value C.

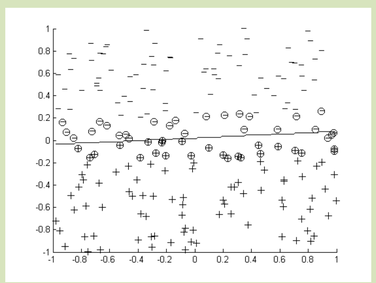

Fig. 1. SVM on the dataset with imbalance ratio 1:1

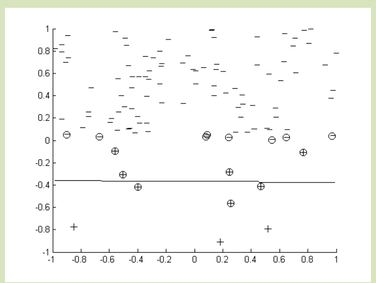

Fig. 2. SVM on the dataset with imbalance ratio 10:1

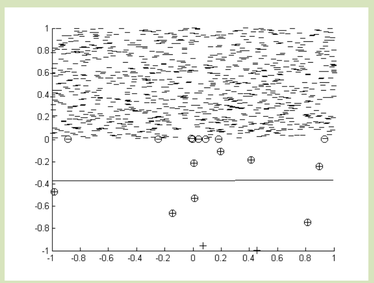

Fig. 3. SVM on the dataset with imbalance ratio 100:1

Methods

Special Case of 2C-SVM

Based on the above analysis, to address the problem caused by $\alpha_i = C$, intuitively, we can set the constraints to be $0 \leq \alpha_i < C$ for the negative class. Recalling the complementary slackness, we have $y_i(\beta^T x_i + \beta_0) \geq 1$, which is equivalent to the hard margin SVM constraint; thus the modification can be further illustrated as the 2C-SVM when .

Constraint on the Slack Variable

Inspired by the fact that misclassification of a positive example usually costs more than that of a negative example, we add one constraint to the slack variables of the positive class to ensure that no positive example is left out and solve the following optimization problem:

Primal:

$$\min_{\beta,\beta_0,\xi_i} \quad \frac{1}{2}\beta^T\beta + C\sum_{i=1}^{n}\xi_i$$

$$s.t. \quad y_i(\beta^T x_i + \beta_0) \geq 1-\xi_i$$

$$\xi_i \geq 0 \quad \text{for} \quad i \in I^-$$

$$0 \leq \xi_i \leq 1 \quad \text{for} \quad i \in I^+$$

* μ is the Langrangian multiplier of the added constraint

Dual:

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) - \sum_{i=1}^{n}\alpha_i + \sum_{i\in I^+}\mu_i$$

$$s.t. \quad \sum_{i=1}^{n}\alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C \quad \text{for} \quad i \in I^-$$

$$\left.\begin{array}{l}\alpha_i \geq 0\\ \mu_i \geq 0\\ \alpha_i - \mu_i \leq 0\end{array}\right\} \quad \text{for} \quad i \in I^+$$

Results

In the experiment, we compare the performance of the proposed approach with regular SVM, 2C-SVM, and undersampling and oversampling techniques.

Table. 1. Testing UCI datasets

Dataset	Imbalanced Rate	Training Set	Test Set
Glass(5)	6.07%	108	106
Abalone(4)	1.36%	417	3760
Car	3.76%	345	1383
Segment(1)	14.29%	231	2079
Segment(3)	14.29%	231	2079
Yeast(ME2)	3.44%	297	1187
Yeast(ME1)	2.96%	297	1187

Table. 2. Experiment Results on 5 UCI datasets

Dataset		SVM	2C-SVM	SVM($C_+ = 0$)	SVM($\xi_+ \leq \theta$)	SMOTE	Undersmp
Glass(5)	F-measure	0.6667	0.5714	0.6667	**0.8333**	**0.8333**	0.6111
	G-mean	0.7071	0.8002	**0.9674**	0.9083	0.8452	0.9225
	AUR	0.7500	0.8135	**0.9654**	0.9117	0.8571	0.9224
Abolone(4)	F-measure	0.3516	**0.4364**	0.4221	0.4210	0.3035	0.2750
	G-mean	0.5330	0.9451	0.8944	0.8653	0.9503	**0.9539**
	AUR	0.6507	0.9454	0.8975	0.8714	0.9503	**0.9542**
Car	F-measure	0.9039	**0.9541**	0.7647	0.9039	0.7161	0.6125
	G-mean	0.9489	0.9533	**0.9879**	0.9489	0.9845	0.9748
	AUR	0.9500	**0.9981**	0.9880	0.9500	0.9846	0.9751
Segment(1)	F-measure	**0.9898**	**0.9898**	0.9882	**0.9898**	0.9640	0.9370
	G-mean	**0.9940**	0.9927	0.9924	0.9927	0.9713	0.9792
	AUR	0.9907	**0.9927**	0.9924	**0.9927**	0.9717	0.9792
Segment(3)	F-measure	0.8493	0.8635	**0.8663**	0.8653	0.7702	0.7754
	G-mean	0.9239	0.9405	**0.9585**	0.9422	0.9080	0.9425
	AUR	0.9248	0.9408	**0.9585**	0.9425	0.9083	0.9430
Yeast(ME2)	F-measure	0.2319	0.3529	0.2963	**0.3546**	0.3545	0.2791
	G-mean	0.4378	0.7410	0.7688	0.7549	0.8111	**0.8256**
	AUR	0.5888	0.7617	0.7892	7722	8172	**0.8264**
Yeast(ME1)	F-measure	**0.6364**	0.5833	0.5667	0.6153	0.5909	0.5246
	G-mean	0.7712	0.9781	0.9635	**0.9803**	0.9670	0.9584
	AUR	0.7957	0.9783	0.9636	**0.9805**	0.9671	0.9585

The results show that SVM (C_+ = 0) achieves high score on G-mean and ROC-curve but is lower than the other algorithms in precision which results in a low F-measure. This can be explained by the hard constraints C_+=0. While the recall rate is guaranteed by including most of the positive examples, it will inevitably include more negative examples. The second approach has a relatively better and more stable performance over all three metrics. In addition, although 2C-SVM has been proposed for quite a long time, it is not given much attention in previous work for addressing the imbalance problem. In our experiments 2C-SVM demonstrates decent performance. As for the re-sampling techniques, SMOTE has better overall performance than random undersampling, but both techniques have low F-measure scores.

Conclusion

In this paper, we propose two modifications of Support Vector Machines to address the problem of classifying highly imbalanced datasets. We study their behaviour comprehensively using three comparison methods. The results show that the two proposed methods have a consistent improvement over SVM's performance. According to our experiments, 2C-SVM is comparable to other algorithms as well, which is neglected in the former studies. The re-sampling methods can mainly be criticized for changing the dataset and introducing unnecessary noise. We conclude that the proposed approaches are promising candidates for addressing the rare class problem.

Reference

[1] N. V. Chawla, K.W. Bowyer, L. O. Hall, andW. P. Kegelmeyer, “Smote: Synthetic minority over sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[2] E. E. Osuna, R. Freund, and F. Girosi, “Support vector machines: Training and applications,” Massachusetts Institute of Technology, Tech. Rep., 1997.